IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 18, 2017, accepted January 15, 2018, date of publication January 25, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2797879

Human-Guided Evolutionary Story Narration

KUN WANG[1], VINH BUI[2], ELENI PETRAKI[3], AND HUSSEIN A. ABBASS[2], (Senior Member, IEEE)

[1]College of Engineering, Ocean University of China, Qingdao 266100, China
[2]School of Engineering and IT, University of New South Wales, Canberra, ACT 2600, Australia
[3]Faculty of Arts and Design, University of Canberra, Canberra, ACT 2601, Australia

Corresponding author: Kun Wang (kunwang@ouc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61601428, in part by the ARC Discovery Program under Grant DP140102590, and in part by the Shandong Postdoctoral Innovation Foundation under Grant 140880.

ABSTRACT Stories are useful tools with which we can exchange experience learnt in social contexts, ways to communicate futures in strategic planning, and unique building blocks that connect meanings in a movie or a virtual environment. Evolutionary computation (EC) techniques have the potential to overcome existing limitations in automated storytelling, whereby evolution can provide a process of innovation. However, one source of complexity lies in the transformation of a story in a natural language into a representation that EC can evolve easily. Another complexity arises from the fact that the ultimate judge for the quality of a story is a human being, and humans are diverse in their taste. This paper attempts to tackle the above complexities through an automatic story narration application. We present a methodology which can transform a story written in English into an event-level and hierarchical-level grammar using a network representation. This approach makes it possible to devise an encoding scheme that translates a story narration with flashback into a chromosome and vice versa. We then discuss different metrics for the evolutionary narration problem and use 42 human participants to evaluate the generated narrations. To incorporate diversified human opinions, we propose to build individual human-surrogate models from the human-evaluation experiment and further fuse them into an ensemble. The ensembles of human surrogate models serve as the objective functions of multi-objective EC to guide the generation of desirable stories from human perspectives. We demonstrate that this approach is successful in evolving better narrations as assessed by 31 human participants.

INDEX TERMS Evolutionary computation, human factors, strategic planning, evolutionary story telling, representation, human–guided evolutionary computation.

I. INTRODUCTION

Stories are means with which we exchange experiences [1]–[3] and the unique building blocks that connect meanings in a video game [4], movie [5] or a virtual environment that produces story-like scenarios for training [6]. Moreover, futuristic stories can capture uncertainty in the world of socio-technical and social systems, such as strategic planning [7], [8]. Computational and automatic storytelling or story generation is becoming a highly intriguing and challenging subject that combines narrative theory[1] with linguistics, psychology, artificial intelligence (AI), and computer science.

Building automatic story generation systems needs to address the following research questions:

(1) Of what is a story composed? From a linguistic point of view, as it is a story we attempt to generate, a story structure — which can indicate the basic building blocks of a story and the relationships or constraints that connect them — is required so that stories based on it can form a coherent whole.

(2) How can a story be computationally represented to facilitate automatic story generation? From a computer science perspective, a computer requires a formalism to represent story structure and the generated stories.

(3) How can a story be evaluated; in particular, how can its 'qualitativeness' be quantified using a computational

[1]In narrative theory, narrative is the umbrella term of story and narration. Most narratologists agree that a story denotes the underlying meaning or content of a narrative while narration corresponds to the process of telling. In this paper, we use the term "story" instead of "narrative" — as has been implicitly followed in the computational storytelling field — to denote narrative that is not tied to any particular medium, such as text or video and in English or Chinese.

model of subjective story metrics, such as coherence, novelty and interestingness?

(4) How can interesting stories or stories with desirable features be generated? Some mechanism is required to guide the generation towards interesting stories instead of dull ones.

The significance of this work is twofold. First, it offers an automated methods to generate different narrations of stories as a useful tool for computational linguists, game designers, and other form of users who rely on digital story narration techniques. Second, the idea of story narration is very useful for simulation scenarios as they offer a variety of methods for playing the same simulation from multiple perspectives without changing the underlying set of events in the story.

Existing approaches have made contributions towards answering those questions. The case-based reasoning approach [5], [9] relies on reasoning about existing stories and recombining parts of them to generate a coherent new story. The simulation or planning-based approach [4], [10]–[12] produces a novel story by recording all the events that happened during one simulation run of a virtual story world where story characters pursue their goals. This approach has become the key to realise ''interactivity'' in interactive storytelling [10]. The traditional search approach regards generation of a story as searching through a network [13], a dynamic map [14], or searching guided by heuristics [15]–[19]. Each searching path forms a new story. An extensive review of automatic storytelling approaches can be found in [20].

These approaches possess several limitations, ranging from the possible lack of creativity, coherence, and interestingness [21] to the absence of human evaluation of the story quality. The quality of the generated stories is varied and there is no mechanism which transforms already generated stories into further better stories.

We conjecture that a possible solution to the problems in existing automatic storytelling approaches is to apply evolutionary computation (EC) to evolve a population of stories. EC relies on an implicit self-feedback loop in which stories generated in one iteration contribute to subsequent generations. The process relies on humans to evaluate the stories to guide evolutionary dynamics. Desirable story features observed in the computational storytelling literature include ''semantic and metrical faithfulness'' in poetry [22], ''suspense'' [23], ''surprise'' [12] and ''coherence, creativity and interestingness'' [20], [24]–[27].

However, one source of complexity of the EC-based storytelling approach lies in transforming a story into a representation that EC can evolve easily, that is, devising an encoding scheme that translate a story into a chromosome and vice versa.

Another complexity arises from evaluating a story, that is, devising a fitness function for EC that estimate the contribution of a story in generating future and improved stories. This requires ''understanding'' a story and therefore giving feedback regarding its intrinsic quality using quantitative metrics. Two types of story metrics have been observed—the objective and the subjective metrics. Objective story metrics are usually defined as heuristic functions or process that can automatically calculate the fitness values of a story [22], [23], [28]. For instance, the ''suspense'' metrics in [23] is defined to estimate an action's contribution to generate a suspenseful story from the context. However, the ultimate judge for the quality of a story is still a human-being since objective metrics alone is still far from predicting exactly how a human would interpret a story in terms of this metrics [18]. Stories ''need to be assessed, either singly or in combination, by human readers'' [29]. Therefore, it would be appropriate to further involve human feedback—in the form of subjective metrics—in the evolutionary storytelling process and apply interactive EC (IEC).

Preliminary results based on a pilot study in [24] and [25] show that through an evolutionary process guided by human assessment of stories regarding some desirable story features (e.g., coherence, interestingness, and creativity), the generated stories can be improved in quality to demonstrate these features to some degree. Nonetheless, this pilot study is subject to significant human fatigue problem of IEC: the human evaluator is unable to evaluate stories over a large number of generations of story evolution due to psychological and/or physical exhaustion.

An effective solution to the human fatigue problem is surrogate-assisted IEC [30]–[32]. Recent works of Wang *et al.* [26], [27] demonstrates the effectiveness of a surrogate model in collecting good story narrations with reduced human fatigue through interactive story evaluation and evolution. However, this preliminary implementation reveals a few problems which include subjective extraction of story structure, bias in human subjective evaluation introduced by a weak experimental design, and lack of variances in the participant samples of the human-based experiments thus fails to facilitate the discussion of diversified human taste and opinions. We address the subjective extraction of story structure problem in a recent work in 2016 [20], while failing to solve the above problems in human evaluation and experiments.

This paper extends existing work on a human-guided evolutionary storytelling approach and attempts to further tackle the above two complexities of EC-based storytelling approach through an automatic story narration application. The story narration problem is a novel contribution in automatic story generation and has wide applications in related planning and scenario generation fields. The system block diagram is presented in Fig.1.

Story narration is the mechanism whereby the same story can paint different mental pictures in a recipient's mind, by carefully manipulating the sequence of events in a story to generate coherent, but different, logical causal inferences. Once essential information of events has been extracted from a story, this problem allows us to explore different ways of manipulating sequence of events, build long-distance causal relationships (e.g. temporal, spatial or interpersonal

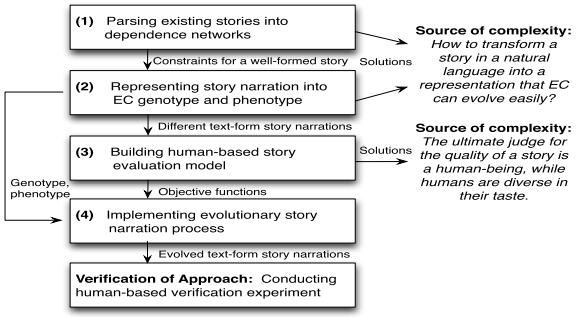

FIGURE 1. System block diagram: (1) to (4) denote our answers to the four research questions of automatic story generation, a link is annotated by the output of a step that will feed to the next directed step.

relationships) and explore possible effect of them.

The first difficulty of the human-guided evolutionary story narration approach is solved by encoding a story narration into a simple linear permutation in this paper. However, as has been pointed out in our recent work [20], the simplicity of the representation comes with the cost of designing a set of linguistic constraints and transformations to guarantee that any random chromosome can get transformed into a unique coherent and causally consistent story narration. This paper addresses the above question (1) of automatic story generation and extends previous work by further extracting the linguistic constraints in the form of a ''hierarchical dependence network'' — not restricted to the event-level constrains — from an existing English text-form story using our revised story parsing method in Section II. A story narration problem can then be transformed into a permutation problem in Section III, which proposes a story representation to answer the above question (2). The extracted dependence network provides the constraints for a valid genotype and guides the decoder for transforming a permutation genome into a valid story narration. This approach makes the evolution of a story possible using classical EC and the many–to–one genotype–phenotype mapping eases the way for efficient evolutionary neutral paths.

Moreover, we undertake a first step towards tackling the second difficulty of devising a human-based story evaluation schema that can incorporate diversified human tastes. Different metrics for the evolutionary narration problem are presented in Section IV. We invited a large sample of human participants to evaluate the generated narrations in a human-based evaluation experiment discussed in Section V. To incorporate diversified human opinions, we propose to build individual human surrogate models from the human evaluation experiment and further fuse them into an ensemble in Section VI. The ensembles of human surrogate models serve as the objective functions (i.e., fitness function) of multi-objective EC to guide the generation of desirable story narrations from human perspectives as well as a computational story evaluation scheme that can incorporate diverse human tastes to answer question (3) of automatic story generation.

A human-guided evolutionary story narration process is presented in Section VII by synthesising the previous sections. A multi-objective evolutionary process evolve a population of story narrations guided by surrogate models of human towards interesting stories, which tackles the above question (4) of automatic story generation.

Verification of the proposed approach in this paper is achieved through a human-based verification experiment discussed in Section VIII. This experiment involves a larger sample of people from different backgrounds so that the findings from the collected data can establish confidence in the approach's capability to handle variations among human participants.

II. PARSING STORIES INTO HIERARCHICAL DEPENDENCE NETWORKS

A story parsing method is proposed to address the first difficulty of the human-guided evolutionary story narration approach. A linguistic approach is applied to firstly transform a story written in English into an event-level grammar using a network representation, which has been elaborated in our recent work [20] and will be briefly discussed in this section. A hierarchical dependence network is further built from the extracted event-level network to facilitate storytelling above the event level.

A. EVENT-LEVEL DEPENDENCE NETWORK EXTRACTION

As there is a strong consensus among narratologists that a story is represented as a sequence of events [33], the basic building blocks (the nodes) of a story dependence network are defined as events.

Essential information of events is also extracted and represented in parameters to facilitate story narration by manipulating the sequence of events in different ways. The parameters of events include the participants (i.e., the characters and objects involved) of events and the temporal and spatial information.

The extracted event parameters serve as clues for building the dependent relations (the links) between events.

1) DEFINITION OF EVENT

In Oxford dictionary [34], event is defined as ''thing that happens, especially something important''. This definition has been further enhanced by TimeML guideline — an international cross-language ISO standard for annotating events from text [35]. The TimeML guideline defines an event as ''a cover term for situations that happen, occur, hold, or take place'' which ''can be punctual or last for a period of time'' and also includes ''those predicates describing states or circumstances in which something obtains or holds true''.

However, the TimeML event definition requires context information to recognise an event, thus, makes it difficult to give an unambiguous event definition for story parsing because some differences between an event and non-event are so delicate that even humans fail to reach a consensus [36]. As our objective is to extract essential event-level information

from a story to facilitate further story narration, it is pragmatic to relax the constraints in the TimeML event definition.

The event definition applied in this paper is: an event is a predicate that denotes an action, state, or occurrence in a story; it is bounded by a position in the temporal dimension, possesses a spatial situation in the story world and has participants as parameters.

2) RECOGNITION OF EVENT

The main part of event in English is represented by a verb because a verb ''forms the main part of the predicate of a sentence'' [37] and ''indicates an action, an event or a state'' [34]. To extract the events from an English text-form story, we need to trace the verbs, or the verb phrases in each of the clauses in an English text-form story. We revise the TimeML event annotation guidelines [35] and extract the following grammatical components in a clause as events.

(E1) VERB without predicative complement
(E2) Predicative complement with NP as the head
(E3) Predicative complement with ADJECTIVE as the head
(E4) Predicative complement with PP as the head

3) DEFINITIONS OF EVENT PARAMETERS

An ontology of event parameters is proposed in this section to extract essential information related to an event, which covers ''who were involved in an event'', ''when and where did an event happen'' to support our computational story narration.

Time is a temporal expression that denotes when the event happened. Three types of time properties are distinguished: the one that directly refers to a position, or situation which means a duration or frequency, in the temporal dimension; the one with reference to another event which builds up a temporal relation between the two events; and blank which means ''in the same time period or close to to the last explicitly mentioned time'' or this time is understandable or not important in the story.

Space is defined as the spatial expression denoting where this story happened in which a concrete thing is involved, such as ''the forest'' in ''in the forest''. Three types of space properties are differentiated: the one with reference to another concrete thing in the story world; the one with reference to another event which builds up a spacial relation between the two events; or blank which means ''in the same space area to the last explicitly mentioned space'', ''close to the last explicitly mentioned space'' or the space of this event is not important in the story.

Character is defined as an active participant[2] that can perform actions to change the states of the story world. A character is a concrete thing that has ever served as the subject of the action verbs, such as ''Ook'' in ''Ook refused to work.'', or a concrete thing that indicates pocessive relationship in the story, such as ''Ook'' in ''Ook's sister''.

Description is defined as a description of a concrete thing in the story world that indicate an participant's identity or state. For instance, ''the most naughty boy'' in ''Ook was the most naughty boy.'', and ''slaves'' in ''Chief made all the villagers work as slaves.''.

Topic is defined as an abstract thing or concept mentioned in the story which can be represented by a sequence of events in the context. For instance, an investment plan, a good idea, a secret, and old happy life.

Object is defined as a concrete thing in the story world that is not a character, a description, a concrete thing in the space parameter, or a time parameter. Compared with a character, an object has not shown as an active participant of the story that can perform actions to change the states of the story world.

We must clarify that this ontology is subject to revision in which more information in a clause can be incorporated to accommodate different storytelling objectives.

4) RECOGNITION OF EVENT PARAMETERS

The event parameters defined in the above ontology can be identified from an English text-form story as follows:

''Time'' and ''Space'' parameters are denoted by the noun phrases (NPs), prepositional phrases (PPs), adverbial phrases (ADVs) and subordinating clauses (SBARs) listed in [38].

''Character'', ''Description'', ''Topic'' and ''Object'' parameters are denoted by NPs in English grammar [38, Nos. 44 and 54]. The following aspects need to be considered during their recognition: firstly, we need to differentiate NPs that denote participants from those that denote time and space information, such as Sunday and 50 meters; secondly, we need a process of coreference resolution, or a Character parameter may not be identified merely because it does not appear in the same label in the story, for instance ''Ook'' may appear as ''Lily's brother'' or ''he''; finally, we need a global view of the story — being a non-actor in one event does not change a participant's identity as a Character parameter if it performed an action in another event.

We further manually combine an event with event(s) that serve as its grammatical components to make its meaning complete and compact. An event can be combined with the events that serve as its: grammatical subject, object, complement and appositive, modifier [38, No. 645]; time-denoting adverbial [38, Nos. 145, 150, 154, and 155]; space-denoting adverbial [38, Nos.161–191]; and direct cause or effect [38, No. 515].

After a story has been separated into parameterised events, some events may be missing the time and space background information due to the strategy to reduce redundancy in natural language. This situation may bring about the difficulty for the reader to make sense of the whole story when we produce new story narrations by shuffling the extracted events or event

[2]In this paper, ''participants'' are used to include not only the traditional ''characters'' in automatic storytelling which usually mean active participants of events that can initiatively act to move the story forward, but also the traditional ''objects'' which are usually inactive thus can only affect the story through interaction with the ''characters'', e.g., being used by a character, triggering a character's attention thus affecting the character emotionally and cognitively.

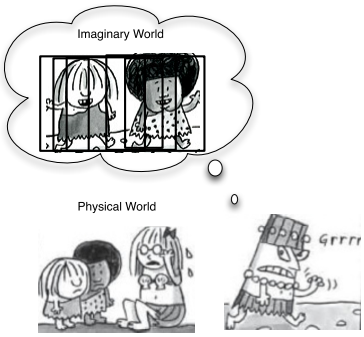

FIGURE 2. Different worlds in CaveLand story.

groups. Therefore, it is necessary to add the implicit time and space parameters — which are basically the time and space parameters in the nearest already-occured event — to the events to complete the missing background information. However, this is not a trivial task. For instance, events may happen in different worlds, as depicted in Fig.2. Therefore, we have only implemented a manual method at this stage.

5) DEPENDENT RELATION BUILDING

A dependent relation is defined on a pair of events in which one event serves as one of the enabling conditions of the other. It can be identified from a counter-factual test [39] which has this form: ''if event A had not happened in the circumstances of the story, then event B would not have happened''.

We propose general rules to manually extract dependent relations between events in a story, in which the extracted event parameters serve as the clues. Event 2 is dependent on event 1 if all of the following conditions are fulfilled:

(R1) Event 1 is one of the nearest events that occurred before event 2 in the story with a shared participant.

(R2) Event 1 must have happened prior to event 2 in the temporal dimension of the story.

(R3) In event 2, the shared participant with event 1 should be aware of the happening of event 1, which usually requires event 1 and event 2 to happen in the same world, either in the objective physical world, or the cognitive world of a participant, etc.

However, it is still a non-trivial task to automate this process considering the challenges in automatic time reasoning [40] and solving the different world problem depicted in Fig.2.

6) PRELIMINARY EVENT GROUPING

We further combine an event with the events that serve as its grammatical subject, object, complement, appositive, modifier and certain types of adverbials to make the meaning of the event complete and compact. The following set of preliminary event grouping rules are proposed.

(G1) Combine the main clause with the nominal subclauses which include that-clause, interrogative clause, to-infinitive and -ing clauses.

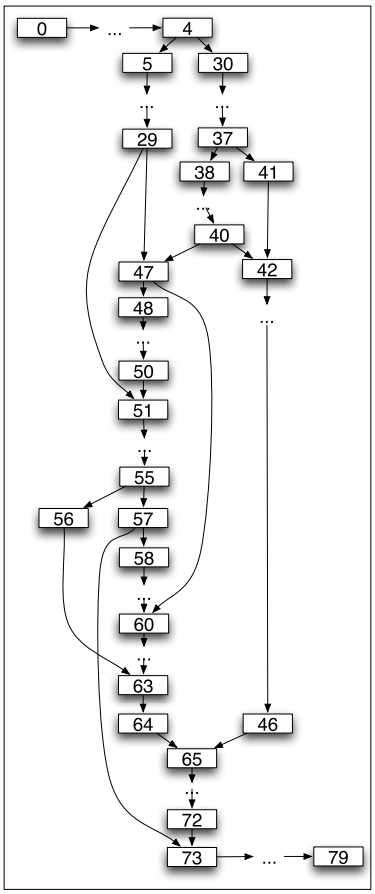

FIGURE 3. An example of story dependence network extracted from "CaveLand" story: the nodes labeled by integers denote the events labeled by their occurrence order in the story; each directed line denotes a dependent relation from one of the enabling conditions of an event to the event; and each node labeled by "..." denotes a chain of events without any branches.

(G2) Combine the main clause with the relative sub-clause.

(G3) Combine the main clause with the time-denoting adverbial sub-clauses for extracting the Time Parameter of this event.

(G4) Combine the main clause with the space-denoting adverbial sub-clauses for extracting the Space Parameter of this event.

(G5) Combine the main clause with non-finite adverbial sub-clauses including -ing clauses, -ed clauses and to-infinitive clauses.

An automatic implementation of the above rules can be facilitated by using a English parser such as the Stanford Parser [41] to obtain the grammatical structure of the sentence then identify those sub-clauses listed in the rules.

Fig.3 illustrates an extracted example dependence network.

B. COMPUTATIONAL REPRESENTATION OF DEPENDENCE NETWORK

All the extracted event parameters can be computationally represented. We highlight event information in terms of the

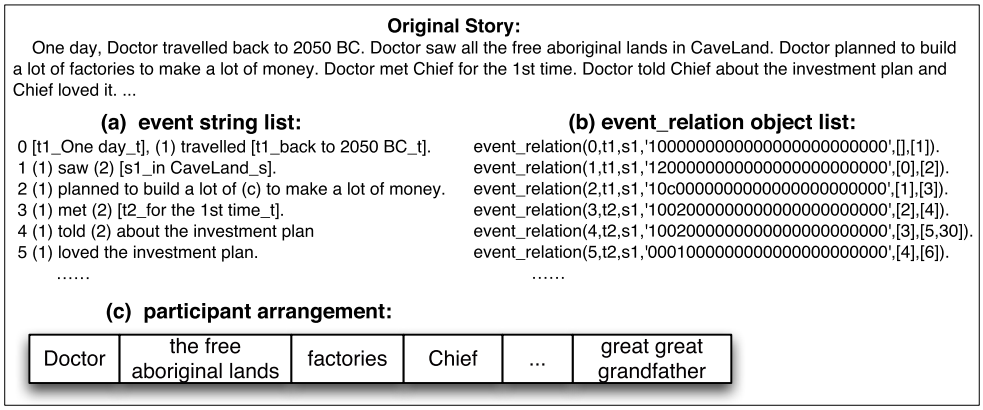

FIGURE 4. An example of computational representation of event-level dependence network: in the "event string" list, (1) denotes that the participant is the "subject" of the event, (2) for "direct object", and (c) for "direct object in the sub-clause".

participants (i.e. the characters and objects) involved, which is different from the time and space information highlighted in [20]. This is due to the different story domains of the two works: this paper focuses on the narration of children's stories, while the previous work generates strategic scenarios in defence and security where "a practical scenario may not require a character at all" [8] or it would not produce coherent and causally-related events by manipulating the characters.

The essential data structures designed to computationally represent the event-level story dependence network in this paper are "event string", "event_relation" and "participant arrangement".

1) EVENT INFORMATION

The event information is described in the "event string", "participant arrangement" structures and the ID, TIME, SPACE and PARTICIPANTS members in the "event_relation" structure. We have not manipulated the Topic, Time and Space parameters of events in this paper. The original text-forms of the Topic parameters are maintained in the "event string". Those of the Time and Space parameters are annotated with labels in the "event string", with the corresponding labels represented in the "event_relation" structure.

a: EVENT INFORMATION IN "EVENT STRING" STRUCTURE

The English text-form of an event is represented. The Time and Space parameters are annotated by a unique symbol ("t" and "s" followed by the their occurrence order in the story) bracketed in "[]". Each participant of the event is substituted by a unique symbol bracketed in "()" in order to facilitate computational story narration in manipulating the participant arrangement of the story, with denotations from a syntactical perspective of English applied. Fig.4 illustrates the original text-form and the extracted "event string" list from the CaveLand story.

b: EVENT INFORMATION IN "PARTICIPANT ARRANGEMENT" STRUCTURE

A permutation of the participants in text-form in the story is used to represent the participants of events. It can be obtained by firstly enumerating all the extracted participants in the story and, then, adding them to a permutation structure.

c: EVENT INFORMATION IN "EVENT_RELATION" STRUCTURE

Every instance of the "event_relation" structure is based on an event in the story. Four members represent the information of an event: ID is the label of an event's identity which is assigned by the event No. — the occurrence order (beginning from 0) of the event in the story; TIME is the a label of its Time parameter which is assigned by a label "t" followed by the occurrence order (beginning from 1) of the temporal expression; SPACE is a label of its Space parameter which is assigned by a label "s" followed by the occurrence order (beginning from 1) of the spatial expression; and PARTICIPANTS is a fixed length string, with each subordinating character denoting the grammatical role that a certain participant played in this event.

The PARTICIPANTS member of the "event_relation" structure can be obtained by replacing each participant's name with a symbol that denotes its participating situation in this event using the following rules: if the enumerated participant has NOT been observed in the event, fill the corresponding position of the enumerated participant in the PARTICIPANTS string with "0"; else, fill the position for the enumerated participant in the PARTICIPANTS string with the corresponding symbol in the bracket ("()" or "{}") in the "event string" instance of this event. Fig.5 provides an example.

2) DEPENDENT RELATION INFORMATION

The "event_relation" structure includes two members that represent the information of dependent relations in the story:

Doctor	the free aboriginal lands	factories	...	great great grandfather
1	2	0	...	0

PARTICIPANTS `12000000000000000000000000'

FIGURE 5. An example of obtaining event No.1's PARTICIPANTS string from "event_relation" structure.

FIGURE 6. An example instance of "event_relation" data structure: two dependent relations are expressed — the one from event No. 0 to 1, described in PARENTS; and the one from event No.1 to 2, described in KIDS.

PARENTS and KIDS which incorporates the IDs of the events that enabled this event and were enabled by this event, respectively. A dependent relation is expressed from the basis event of an "event_relation" instance to each of the events whose ID have been included in the KIDS member, or from each of the events whose ID have been included in the PARENTS member to the basis event. Fig.6 illustrates an example "event_relation" instance.

C. HIERARCHICAL DEPENDENCE NETWORK BUILDING

Some events in the dependence network are so closely related that any interruption during the narration may cause the plot to break into small pieces thus should be avoided. These close dependent relations exist in a chain of reactions to an event by one participant or a constant group of participants, or a chain of interactions between two participants, etc. So we need to group the closely related events together.

We can observe macro story building blocks in computational storytelling and narrative theory, such as "functions" [5], [42], "scenes" [10] or "episodes" [43], which can be described by a sequence of events. However, a consensus about the definition of the macro story building blocks is unavailable. We believe that a pragmatic solution is to admit a possible hierarchy, such as the beats, shots, scenes and acts hierarchy [16] in the movie production industry.

Some consistent indicators of macro story building blocks still exist, such as change of space, explicit reference of time, participating characters or objects. In [44], each scene corresponds to the events that took place at a specific locale. In [43], some basic units are clustered by all viewers including contiguous shots of the same "setting" which specifies the locations, objects and characters that are present in the scene.

Our proposed parameterised event-level dependence network can serve as a pragmatic model that facilitates the unambiguous, computational and hierarchical grouping of events.

The following rule (G6) to (G10) are designed to group the events in a story into a hierarchy of macro building blocks, which refer to as "chains" in this paper.

A new chain can be identified when the following situations are encountered: a branching point and a meeting point in the event-level dependence network for rule (G6) and (G7), respectively; a change of time reference for rule (G8), a change of space reference for rule (G9); and a change of participants for rule (G10).

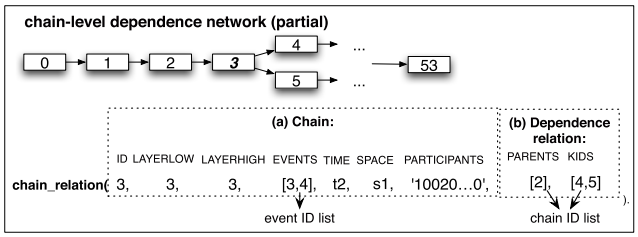

FIGURE 7. An example instance of "chain_relation" data structure.

Rule (G6) and (G7) are firstly applied based on the dependent relation information between events, followed by a flexible combinations of rule (G8), (G9) and (G10) based on the event information in the event parameters. Applying different combinations of rule (G8), (G9) and (G10) will produce story dependence networks that are defined on diversified macro levels. In this paper, we apply rule (G8), (G9) and (G10) to obtain a "chain-level" dependence network. Figure 7 shows an example of chain-level story dependence network extracted from "CaveLand" story. These rules are explained below.

(G6) Group a sequence of dependent events that have enabled the happening of two or more than two independent events, i.e., events that have no dependent relation in between. This is implemented by always starting a new "chain" on each of the events whose ID has been included in the multiple KIDS members of the corresponding "event_relation" instance of the current event, which means a branching point is encountered during the traveling through the story dependence network from the enabling events to the enabled events.

(G7) Group a sequence of dependent events that have been enabled by two or more than two independent events. This is implemented by always starting a new "chain" whenever the corresponding "event_relation" instance of an event have more than one PARENTS members, which means a meeting point is encountered during the traveling through the story dependence network from the enabling events to the enabled events.

(G8) After applying rule (G6) and (G7), further group a sequence of dependent events that have the same Time parameter. This is implemented by always starting a new "chain" whenever the value of the TIME member of the corresponding "event_relation" instance of the event is changed during the traveling through the story dependence network from the enabling events to the enabled events.

(G9) After applying rule (G6) and (G7), further group a sequence of dependent events that have the same Space parameter. This is implemented by always starting a new "chain" whenever the value of the SPACE member of the

corresponding ''event_relation'' instance of the event is changed during the traveling through the story dependence network from the enabling events to the enabled events.

(G10) After applying rule (G6) and (G7), further group a sequence of dependent events that contain the same group of Character and Object parameters, which means always start a new ''chain'' whenever one or more than one new participants have been observed in the current event during the traveling through the story dependence network from the enabling events to the enabled events. This is implemented by always starting a new ''chain'' whenever the below situation happens: in the PARTICIPANTS string in the ''event_relation'' instance of the current event, the value of one character is neither 'D' nor '0' while the value of the character at the same position has always been observed as 'D' or '0' in any of the PARTICIPANTS strings of the previous events in the newly obtained chain so far during the traveling through the story dependence network from the enabling events to the enabled events.

Prolog is implemented to automatically group events. Each of the event_relation structure instances is bracketed in a predicate called ''event_relation'' to serve as the input facts of Prolog. The Prolog program can output a list of ''chain_relation'' predicates whose parameters serve as the values of TIME, SPACE, PARTICIPANTS and EVENTS members of the ''chain_relation'' data structure object discussed. Then the Prolog output is processed by a Java program to assign the value of the other members of the ''chain_relation'' structure.

A ''chain_relation'' data structure (illustrated in Fig. 7) is introduced to computationally represent the chain-level story dependence network. It has a similar construct to the ''event_relation'' data structure except for two members: EVENTS is an ordered list of its contained events' ID where dependent relations between two events are represented in each pair of the two adjacent event IDs; LAYERLOW and LAYERHIGH denote a chain's flexible layers — a layer range — in the story dependence network which is important to generate story narrations with both forward narration and flashback. Fig.8 illustrates the computational representation of a chain-level story dependence network.

This hierarchical story dependence network provides the constraints that define a well-formed story in our domain of interest and facilitates story narration on multiple macro levels above the event level. This hierarchical network serves as our story structure and answers question (1) of automatic story generation.

III. STORY NARRATION AS PERMUTATION PROBLEM

A story narration is represented into EC genotype and phenotype in this section, in which a linear story representation is proposed to answer question (2) of automatic story generation.

We represent a story narration with ''flashback'' as a permutation, that is, finding an appropriate permutation of the chains (or events) and participants in the story dependence network, whereby evolutionary story narration is similar to a traveling salesman problem. The main novelty is lying in the genotype–to–phenotype mapping which can transform a classical permutation genome into a valid story narration. The dependence network extracted in the last section provides the constraints for a valid genotype and guides the decoder for this transformation.

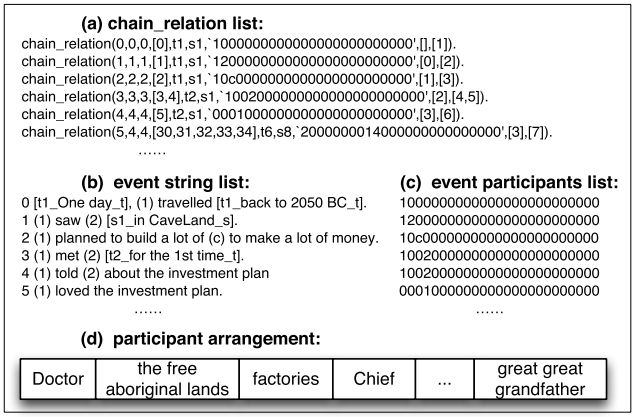

FIGURE 8. An example of chain-level story dependence network.

We incorporate flashbacks in the generated story narrations to manipulate the sequence of event chains thus the reader's understanding of the story. A flashback is a story played backward. It can be a full flashback in which the whole story is played backward, or a partial flashback in which a subset of chains are played backward, with the rest of the chains played forward. Flashbacks ''have much to do with memory'' and have a role of ''guiding the viewers' comprehension of events'' [19], ''enlightening, haunting, surprising, and changing our beliefs towards story events'' [45].

A. ENCODING STORY NARRATION INTO GENOME

As it is necessary to control the coherence in the generated story narrations for humans to provide a comparatively objective evaluation, the following constraints are imposed: a story is told by combining both a forward narration and flashback of the chains in the story dependence network.

1) CONSTRAINTS FOR STORY NARRATION

The above constraints can be realised in the following way, which is similar to our previous work [20] but on the chain level of the story dependence network: firstly, randomly choose a layer in the story dependence network as a threshold layer; secondly, all chains with smaller layer values are narrated in the forward direction, which means a chain will only be narrated when all the chains in its PARENTS chains have been narrated; thirdly, chains with bigger layer values are narrated in a flashback way, which means a chain will only be narrated when all the chains in its KIDS chains have been narrated; finally, the narration will end at the chain with the threshold layer value — the meeting point of the forward and flashback narration of the story dependence network.

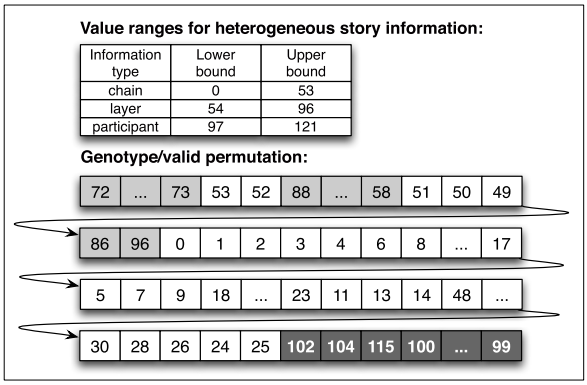

FIGURE 9. An example of genotype: the white genes denote the chains, the highlighted dark grey genes the participants in the story, and the light grey genes the layers in the dependence network.

The diversity of the generated narrations can still be maintained under these constraints because: different layer thresholds will generate narrations that end at diversified points in the dependence network; and the number of combinations of alternations of the forward and flashback chains can be large, as well as the options when parallel paths are encountered during the narration of the chains in the dependence network.

2) INCORPORATING HETEROGENEOUS STORY INFORMATION

The incorporation of heterogeneous story information in a genome is achieved by assigning unique value ranges to different types of information (suppose that there are M event chains and N layers in the story dependence network and the overall number of different participants in the story is P): genes with values between 0 and M−1 denote chains; genes with values between M and M+N−1 denote layers; the first layer gene that occurs in the genotype is the threshold layer while all the following layers' genes are redundant; and genes with values between M+N and M+N+P−1 denote participants in the story, and their order of occurrence in the genotype determines the participant arrangement of the narration.

The genotype can be generated in two steps: firstly, a permutation of M+N+P integers (from 0 to M+N+P−1) is randomly generated which may not meet the constraints mentioned above; and, then, it is transformed into a corresponding permutation that conforms to those constraints. During the transformation process, the dependent relation information represented in the dependence network's chain_relation list — the PARENTS and KIDS members — serves as the reference for checking if the constraints are fulfilled; and the obtained valid permutation serves as the genotype. An example genotype is presented in Fig. 9.

B. OBTAINING TEXT-FORM STORY NARRATION FROM GENOTYPE-PHENOTYPE MAPPING

The text-form story narration can be obtained from a genotype-phenotype mapping procedure in the following steps: firstly, scan the genome from left to right to extract a list of chain genes and participant genes (see Fig.10); secondly, extract the participant arrangement from the participant list; finally, enumerate the text representation of each of the chains one after another in the order of their positions in the chain list. The text of a chain is obtained by enumerating the text representation of each of its contained events in the EVENTS member, as illustrated in Fig.11.

An extra operation is required to adjust the time and space background information of each chain to assist the reader's understanding of the generated narration with flashback. This is performed as follows: if the currently discussed chain possesses the same TIME and SPACE parameters as the previous chain in the genome, neglect the parts annotated by these parameters in the 'event string' of any of its subordinate events during the narration; if the currently discussed chain possesses different TIME and/or SPACE parameters from the previous chain in the genome, narrate the text representation of these different parameters first and neglect narrating the same TIME or SPACE parameters to the previous chain in the genome.

Different text-form story narrations can be generated accordingly and subject to human evaluation in the next stage.

IV. STORY METRICS SELECTION

Evolving story narrations requires the evaluation of them (i.e., defining a fitness function for them). Subjective story metrics for human evaluation are presented in this section. Objective story metrics are defined to reflect the underlying features that are possible to affect a human reader's understanding and evaluation of the narration.

A. SUBJECTIVE STORY METRICS

As it is difficult to quantify the quality of a work of art such as a story [24], a pragmatic solution is to involve humans in the evaluation process, such as asking them to give scores. However, there may exist multiple factors that affect his or her evaluation; for instance, questions such as "is this story easy to understand?", "is this story new to me?", and "has my interest been triggered by this story?". So the story quality is sub-categorised into four different subjective story metrics in this paper. Although further sub-categorisations of the subjective metrics may be possible, a sequence of problems may emerge: it is both time consuming and confusing for a person to give too many scores to a single story narration; and data fusion of all the scores is a challenging task in a multi-criteria decision problem [46].

1) COHERENCE

Coherence reflects "a global representation of story meaning and connectedness, which is the temporal and causal structure of a story" [47] and makes a story understandable to the reader [48].

2) NOVELTY

Novelty reflects the unexpectedness and rule-breaking degree of a story and serves as a supplement to the coherence

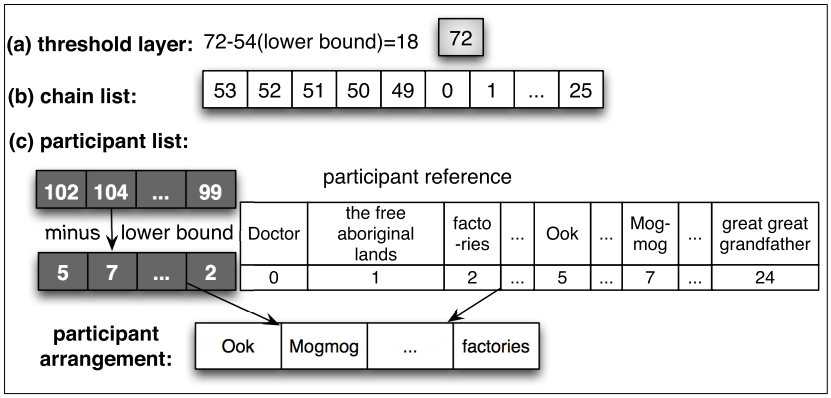

FIGURE 10. Story information extracted from genotype.

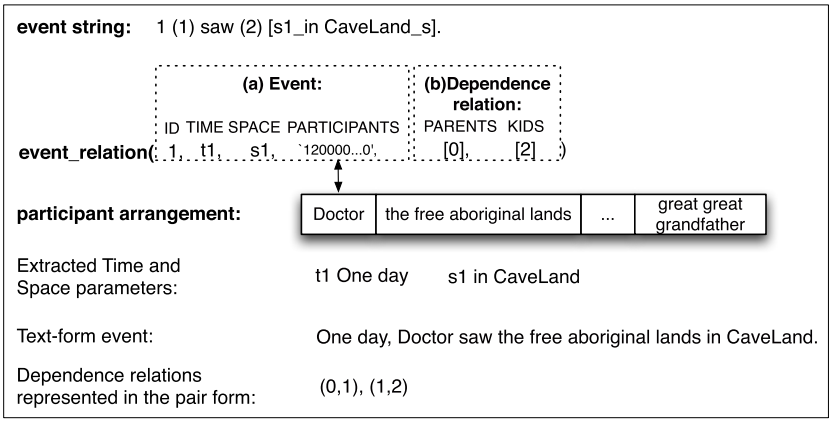

FIGURE 11. Extraction of text representation from sample event.

measure. One objective of computational storytelling is to help discover stories or structures that exceed our imagination so as to achieve some degree of creativity which is a fundamental characteristic of human intelligence and an inescapable challenge for any form of AI [49].

3) INTERESTINGNESS

Generating interesting stories is another general objective of storytelling. If we say that the above two metrics reflect a readers's global impression of a story after understanding is achieved, interestingness may indicate the dynamics of a human's appreciation of a story in the sense that "the increases in cognitive interest were observed before full comprehension was achieved" [50].

4) OVERALL QUALITY

Overall quality indicates a human's overall impression on the quality of a story.

B. OBJECTIVE STORY METRICS

Different degrees of flashback and deviation from the smooth flow of causality represented in the original story may have diversified effects on a human reader's understanding and evaluation of a story. Also, different ways of participant arrangement shuffling may change a reader's mental picture of the story [51] and therefore manipulate his or her understanding of the story to a different degree.

Four types of objective story metrics representing the quantitative features of a story narration are defined and can be generally classified as objective metrics about: the logical structure of events which includes disOfFlashback and consistChainOrder; and the participant arrangement which includes consistParRoles and consistPars.

1) disOfFlashback

The distance of the flashback feature, DOF, of a story narration compared with the original story. Let $DNLN$ be the dependence network layer number and TL the threshold layer, DOF is calculated as

$$DOF = \frac{DNLN - TL}{DNLN} \tag{1}$$

2) consistChainOrder

The consistency of the chain order, CC, of a story narration with the chain order of the original story. Let $SCOSCO$ be the sorting cost to the original story's chain order and n the

number of chains in the dependence network. The sorting cost is calculated using bubble sort.

$$CC = \frac{SCOSCO}{n \times (n-1)/2} \tag{2}$$

3) consistPars

The consistency of the participant arrangement, *CP*, of a story narration with that of the original story. Let *TPCOS* be the times of participant change from the original story and *NPS* number of participants in the story.

$$CP = 1 - \frac{TPCOS}{NPS} \tag{3}$$

4) consistParRoles

The consistency of the arrangement of participants roles, *CPR*, of a story narration with those of the original story. Let *TRCOS* be the number of times of role change from the original story and *NPRS* the number of participant roles in the story.

$$CPR = 1 - \frac{TRCOS}{NPRS} \tag{4}$$

V. HUMAN-BASED EVALUATION EXPERIMENT

A human-based evaluation experiment is conducted to collect values of subjective story metrics from humans. 42[3] human participants are invited to evaluate and assign values of subjective metrics (aka give scores) to 10 selected sample stories with various values of objective metrics. It improves our preliminary implementation [26], [27] in the following aspects: firstly, it fully applies the story parsing approach proposed in Section II; secondly, clear definitions of the subjective metrics are provided to the human participants before story evaluation to train them in giving proper evaluations; and, finally, it involves a larger sample of people from different backgrounds so that the findings from the collected data can establish confidence in the approach's capability to handle variations among human participants, which is also absent in our recent work [20].

A. EXPERIMENTAL DESIGN

1) DEFINITIONS OF SUBJECTIVE METRICS TO HUMAN PARTICIPANTS

The following definitions are printed on a handout provided to each of the human participants before story evaluation.

- **Coherence** denotes a global representation of story meaning and connectedness which is the temporal and causal structure of a story.
- **Novelty** means the way the story organise its characters, time, space and causal relationships, etc., is different, new, unexpected or surprising to you.
- **Interestingness** means you think this story is funny, or your curiosity, expectance, suspense or imagination is triggered when you read the story.
- **Overall quality** means your overall impression on the quality of this story.

2) HUMAN PARTICIPANT SAMPLE

The population from which the sample is drawn consists of 42 human participants, mostly postgraduate students and staffs in the university and a few volunteers recruited from local areas. As most participants (33 out of 42) are aged between 20 to 30, such a sample is skewed to a particular age group. However, a variety of evaluation results is anticipated due to differences in the following individual characteristics.

- **Genders** cover females and males.
- **Language backgrounds** cover native English speakers in local residences, official English speakers (i.e., people who use English as their official languages) and non-native English speakers in international students and staff.
- **Discipline backgrounds** include natural science (e.g., mathematics and physics), engineering and technology (e.g., IT and mechanical engineering), and social science (e.g., business and geography).
- **Working statuses** include students and staff.

3) STORY NARRATION SAMPLE

All the story narrations in the sample are generated based on the dependence network extracted from an existing 'Cave-Land' story which is revised from a recent comic book 'Ook and Gluk' and the traditional 'Little Red Riding Hood' story.

10 story narrations are selected to incorporate big variance in the values of their objective metrics explained in Section IV-B in the following steps: firstly, the value space [0,1] of the objective metrics is divided into two ranges: "LOW" for values in [0,0.5) and "HIGH" for those in [0.5,1]; and, then, 9 story narrations with all the possible value range permutations are selected.

Only 9 (rather than 16) value range permutations of the objective metrics can be observed considering the following inherent relations between the objective metrics.

- **Low values of DOF result in high values of CC.**
- **Values of CPR are higher than those of CP.**

The following value combinations of any two types of objective metrics are possible: "LOW HIGH", "HIGH LOW" and "HIGH HIGH" for DOF and CC; "LOW LOW", "HIGH LOW" and "HIGH HIGH" for CPR and CP. The 10 story narrations in the story narration sample are obtained by firstly randomly generating 500 story narrations using the initial story narration method in Section III and, then, selecting 10 story narrations with the above permutations of objective metrics. Fig. 12 and Table 1 provide extra information of the story narration sample.

The ontology of event parameters provides a natural classification of the participants from which more complicated participant roles can be designed. We define roles by adding derogatory, commendatory or neutral meanings in this paper.

[3]We have recruited 42 human participants for both experiments discussed in this paper. 11 of them apologised and were subsequently absent in the verification experiment discussed in Section VIII.

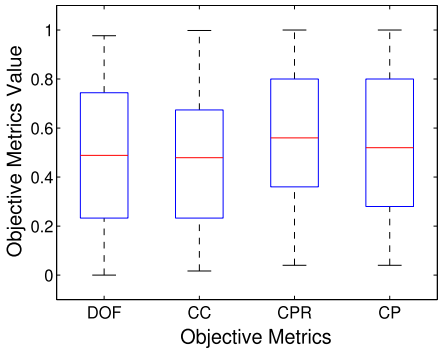

FIGURE 12. Objective metrics distributions of candidate story narrations for story narration sample in human-based evaluation experiment.

TABLE 1. Values of objective metrics in story narration sample in human-based evaluation experiment.

No.	DOF	CC	CPR	CP
1	0.047	0.822	0.320	0.200
2	0.279	0.854	0.560	0.400
3	0.163	0.680	0.680	0.680
4	0.837	0.242	0.200	0.200
5	0.791	0.391	0.560	0.480
6	0.744	0.136	1.000	1.000
7	0.558	0.618	0.120	0.040
8	0.558	0.585	0.560	0.480
9	0.721	0.534	0.880	0.840
10	0	1	1	1

Each human participant is required to give a score (ranging from 0 to 10 in which 0 denoting an extremely undesirable story narration and 10 a great one from the human participant's perspective) to each subjective metrics of a story narration. The definitions of the subjective metrics are provided on a handout before the story reading and evaluation process. The time duration for reading each story for each human participant is also recorded as reference.

B. RESULTS AND ANALYSIS

1) HUMAN PARTICIPANT SAMPLE

42 human participants volunteered to participate in the experiment. The bias in the human participant sample can be observed in the working status and age group distributions: the majority are students (33 out of 42) aged between 20 to 30 (33 out of 42, with 2 aged between 30 to 40, 2 for 40-50 and 4 for 60 and above). The findings from the collected data can still establish confidence in the system's capability to handle variations among human participants considering the variety in gender, language background and discipline in which each group account for at least 30% of the human participant sample.[4]

Distributions of genders in the human participant sample are 20 females versus 22 males. Those of language backgrounds are 13 native English speakers, 24 non-native English speakers and 5 official English speakers. Those of disciplines are 10 with natural science background, 15 in engineering and technology, and 17 in humanity and social science.

Possible degrees of tiredness of human participants are reflected in two factors, "sleep hours before experiments" and "awake time" which denotes the time elapse since the human participants got up. In terms of the former factor, 90.32% (39 out of 42) of the human participants slept from 6.5 to 9 hours before the experiment. Regarding the later factor, 47.6% (20 out of 42) of the human participants had been awake for 3.5 to 6.5 hours before the experiment, 30.95% for 6.5 to 9.5 hours, 19.05% for 9.5 to 12.5 hours, with one for 32 hours.

[4]Although the "official English speaker" group only accounts for the minority of the 42 human participants, it can be merged into either the "native English speaker" or "non-native speaker" group during analysis.

TABLE 2. Correlations between objective and subjective metrics among all human participants in a human-based evaluation experiment.

	DOF	CC	CPR	CP
coherence	-0.495	0.542	0.682	0.685
novelty	-0.479	0.534	0.601	0.619
interestingness	-0.557	0.590	0.608	0.620

2) CORRELATIONS BETWEEN OBJECTIVE AND SUBJECTIVE METRICS

Data of subjective metrics is obtained from the corresponding scores of the 10 sample story narrations provided by the 42 human participants (420 scores in total for each subjective metrics). Table 2 shows the correlations between the objective and subjective metrics among all human participants.

3) RELATIONS BETWEEN SUBJECTIVE METRICS, READING ORDER AND TIME

Figure 13 shows the following interesting relations.

Firstly, in terms of the reading order variable, there is a decreasing trend of the reading time when the reading order increases, which is reasonable considering that human participants develop familiarity with the characters, objects and the events in the story after reading more and more narrations of the story and, thus, gradually digest the story faster. However, variance of coherence, novelty and interestingness values can be observed for a certain reading order, which indicates no significant effect of the reading order on human participants' subjective story evaluation.

Secondly, the majority of the observed time duration for reading one story ranges from around 100 seconds to 300 seconds (around 1 minute and 40 seconds to 5 minutes) while, a variety of coherence, novelty and interestingness values can be observed for a certain reading time.

Thirdly, regarding the subjective metrics, in general, positive linear relations can be observed between any two subjective metrics, especially between the novelty and interestingness in which the variance of the interestingness values for a certain novelty value among different human participants is

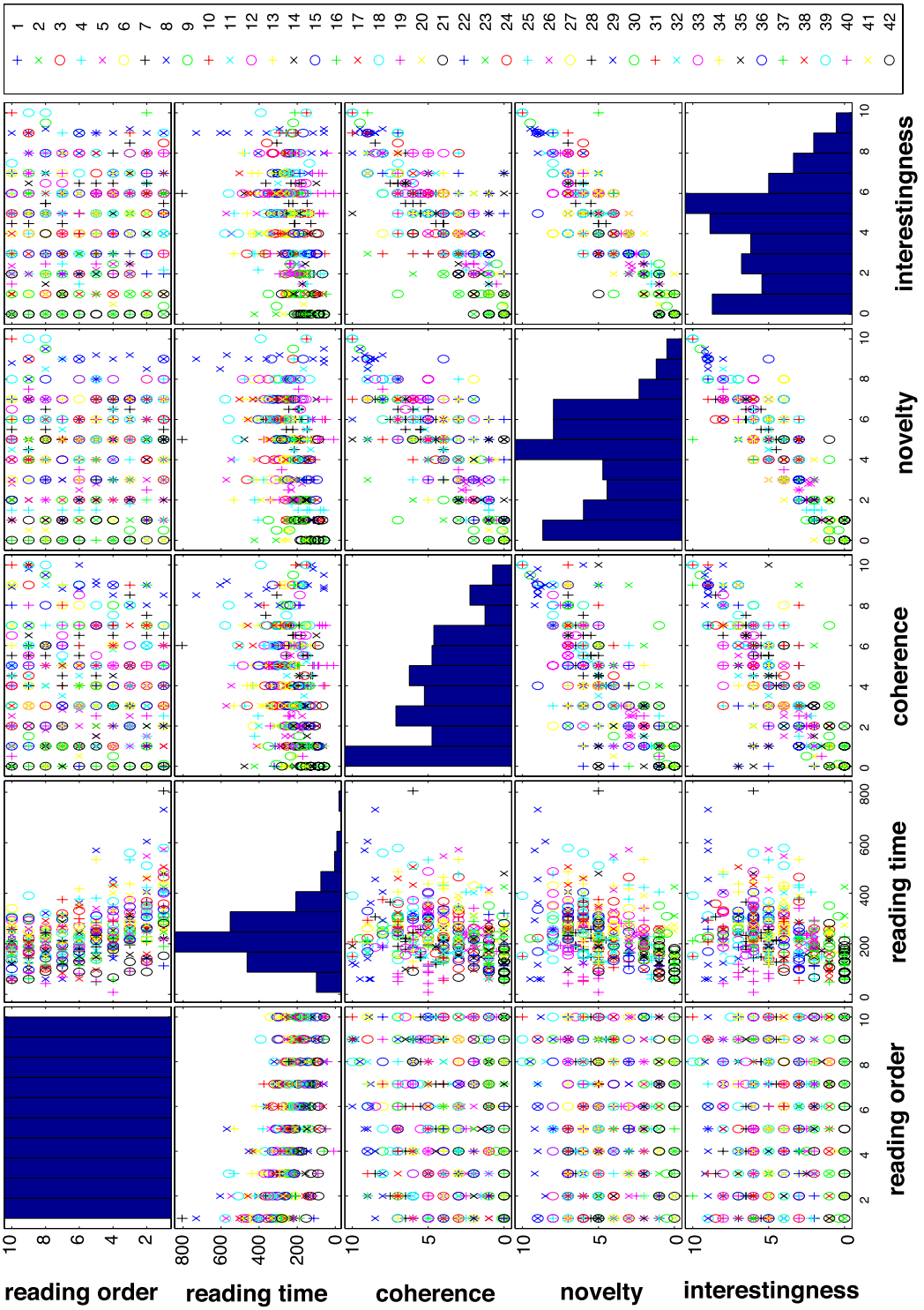

FIGURE 13. Two-two relation plot matrix between observed subjective story metrics, reading order and recorded reading time in human-based evaluation experiment: plots between the same variables compare the occurrence frequency of different values.

smaller than that of the coherence values. This phenomenon implies: human participants may regard novelty as an important factor that contributes to story interestingness; the human participants' evaluation of the story novelty or interestingness may be affected by their evaluation of story coherence, which is reasonable given that, when the coherence is low, people may have difficulty in understanding the story, not to mention appreciating how interesting and novel it is; and, finally,

under the same coherence situation, the human participants' evaluations of novelty or interestingness of story vary, which indicates diversified opinions in what makes a novel or interesting story.

VI. SURROGATE MODEL OF HUMAN STORY EVALUATION

This section builds a surrogate model for human story evaluation based on the subjective story metrics data collected in the above human evaluation experiment. This surrogate model improves our previous works [20], [27] in the way that it can incorporate the diversity of human opinions and has the potential to maintain higher precision in the prediction of human story evaluation.

We propose the use of individual surrogate models, whereby a model is constructed for each human evaluation independently, then the models are weighted with their coefficient of determination and get fused into an ensemble.

These ensembles of human surrogate models can incorporate diverse human tastes and provide a computational story evaluation scheme to answer question (3) of automatic story generation.

A. DATA PREPROCESSING

The collected subjective metrics data is normalised by the values of the tenth story in the story narration sample in Section V-A3 — the original story — using Equation (5) to alleviate possible effect of different value ranges among human participants.

$$Sbj_{norm}[i] = \frac{Sbj[i]}{Sbj[10]} \times 10, \quad i = 1, 2, \ldots, 10 \tag{5}$$

B. INDIVIDUAL SURROGATE MODELS FOR HUMAN STORY EVALUATION

For each human participant, we build a set of individual surrogate models for story evaluation, each captures the mapping between a particular subjective metrics and all the objective metrics of any story narration. At this stage, we apply multiple linear regression using ordinary least squares method [52] represented in Equation (6). The notation is explained as follows: Y is an $n \times 1$ vector representing n cases of observed data about a subjective story metrics which is collected from the above experiment in the form of human evaluation scores for n story narrations in the sample; β is a 5×1 vector of regression coefficients each of which denotes an objective story metrics' weight in determining the value of a subjective story metrics, including the intercept; X is a matrix that gives all the observed values of the objective story metrics; and e is the $n \times 1$ vector of statistical errors.

$$Y = X\beta + e \tag{6}$$

C. FUSING INDIVIDUAL SURROGATE MODELS INTO ENSEMBLE

The individual surrogate models are fused into an ensemble. Individual models with higher R^2 values — higher precision in terms of predicting human story evaluation using the values of objective metrics — have more influence in the ensemble.

TABLE 3. Ensemble of individual linear regression models of subjective metrics as surrogate model.

	c1[1]	c2	c3	c4	intercept[2]
overall	2.038	6.252	-11.960	14.399	0.633
coherence	3.583	9.092	-15.927	18.972	-2.815
novelty	2.603	6.746	-15.024	16.594	1.279
interestingness	2.279	6.833	-13.888	15.844	0.637

[1] c1, c2, c3 and c4 denote the coefficient of the DOF, CC, CPR and CP objective metrics in the linear regression model, respectively.
[2] The intercept in the linear regression model.

Table 3 shows the surrogate model for each subjective metrics. An ensemble is obtained in two steps: firstly, the R^2 values of the N individual models are normalised into 0 to 1 using Equation (7); then, the regression coefficients vector of the ensemble is calculated using Equation (8) in which the normalised R^2 values obtained in the last step determine the individual models' weights in the aggregation of the ensemble.

The notations used in Equation (7) and (8) are explained as follows: $[i]$ denotes the corresponding variable for a particular human participant, $R^2_{norm}[i]$ the normalised R^2 value of human participant No.i, $R^2[i]$ the R^2 value of participant No.i, N the overall number of human participants which is 41 for the overall metrics (one participant's overall scores are not available) and 42 for the other subjective metrics, $\beta[i]$ the β vector of human participant No.i's individual model, $\beta[ensemble]$ the β vector of the ensemble aggregated from the individual models.

$$R^2_{norm}[i] = \frac{R^2[i] - \min_{j=1}^{N}(R^2[j])}{\max_{j=1}^{N}(R^2[j]) - \min_{j=1}^{N}(R^2[j])}, \quad i = 1, 2, \ldots, N \tag{7}$$

$$\beta[ensemble] = \frac{\sum_{i=1}^{N}(R^2_{norm}[i] \cdot \beta[i])}{\sum_{i=1}^{N} R^2_{norm}[i]} \tag{8}$$

VII. MULTI-OBJECTIVE EVOLUTIONARY STORY NARRATION

Story narrations are evolved using a multi-objective evolutionary story narration process in this section, in which human-guided EC guides the generation towards interesting stories to tackle the above question (4) of automatic story generation.

The genotype and phenotype have been discussed in Section III. The surrogate model obtained in the last section automatically assigns fitness values to the generated story narrations. Compared with single-objective story evolution [24], [25], multi-objective evolution has the advantage of maintaining better diversity in the produced story narrations because it holds a pareto-front of solutions based on the objective functions [53].

A. OBJECTIVE FUNCTIONS

We apply an automatic evaluation method using the surrogate model obtained in Section VI rather than a full human-in-the-loop evaluation method used in existing applications [24], [25] to minimise human involvement in evolutionary storytelling.

The coherence, novelty and interestingness subjective metrics are selected as the multiple objectives for the evolutionary process. Corresponding surrogate models are applied to automatically assign fitness values to each generated story narration in the following steps:

Step 1: Extract the heterogeneous story information — the threshold layer, chain list and participant list — from the genotype elaborated in Section III-A;

Step 2: Calculate the objective story metrics values using equation (1), (2), (3) and (4) based on the story information obtained from step 1;

Step 3: Calculate the approximated subjective story metrics values using the surrogate model obtained in Section VI-C based on the objective story metrics values obtained in step 2.

B. ELITISM STRATEGY

The elitism strategy in NSGA-II [54] is applied to maintain elitist solutions in the population during the evolutionary story narration process. Population is sorted based on non-domination. The new generation is filled out by each front in the mixed population of parents and offsprings until the population size reaches a predetermined upper limit. A binary tournament selection based on the crowding distance is used to select parents from the population for crossover and mutation.

C. GENETIC AND SEARCH OPERATORS

- **Crossover operators**: the traditional partially mapped (PMX), order (OX) and cycle (CX) crossovers in [55] are applied, with the pseudo codes introduced in [56];
- **Mutation operators**: the traditional inversion, insertion, displacement and reciprocal exchange operators in [55].

After a mutation or crossover operator is applied to produce offsprings, an extra work is to transform each offspring's genome into one that conforms to the constraints in Section III-A1.

D. EXPERIMENTAL STUDY OF EVOLUTIONARY PROCESS

We test the performance and discuss the effects of the multi-objective evolutionary story narration process in this section.

After testing the performance of the story evolutionary process under different parameter settings, the following parameters are selected. The population size is set to 300 for 5000 generations, 0.8 crossover rate where the PMX, OX and CX crossover operators share equal probabilities, and 0.2 mutation rate where the inversion, insertion, displacement and reciprocal exchange mutation operators share equal probabilities.

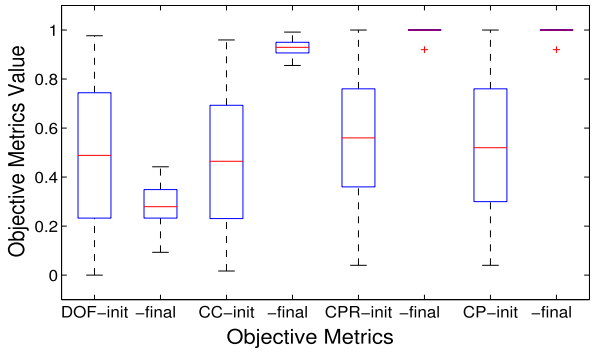

FIGURE 14. Objective metrics distributions of initial and evolved story narrations.

Fig.14 presents a comparison of the distributions of objective metrics in the initial and final population annotated by "-init" and "-final", respectively. On the one hand, from the DOF and CC objective story metrics, the surrogate model for human story evaluation guides the evolutionary process to converge to story narrations whose chain orders are more consistent with those of the original story while still allowing a certain degree of flashback in the evolved story narrations. On the other hand, the evolutionary process tends to leave little diversity for the CPR and CP objective metrics, which indicates that the story narrations obtained through evolution will probably maintain the participant arrangement of the original story.

The transition of each of the evolution objectives — the coherence, novelty and interestingness subjective metrics — during the evolutionary process is illustrated in Fig. 15 and 16. The two figures reveal some interesting insights:

Firstly, the evolutionary process succeeded in collecting story narrations with improved quality in terms of the approximated coherence, novelty and interestingness subjective metrics, reflected in the increasing trend in the plots in the two figures. The average values of all the subjective metrics (i.e. the objective functions) reach 9.0 (90% of the best objective values 10) before generation No. 2000. The subjective metrics values for the best individual in generation No. 5000 are 9.553 for overall, 9.813 for coherence, 9.945 for novelty and 9.689 for interestingness.

Secondly, no obvious conflict between the three objectives is observed considering the constant increasing trend in the plots. A multi-objective evolutionary framework is still a safe choice considering that this situation may change for a different story dataset [20].

VIII. HUMAN-BASED VERIFICATION EXPERIMENT

Before going into the details of the verification experiment, we need to explain why verifying a storytelling system is not a trivial task. The difficulties involved limit our ability to compare many different storytelling methods as compared to computer-based experimentation.

First, scalable test problems similar to those being used in evolutionary multi-objective optimisation [57] are absent

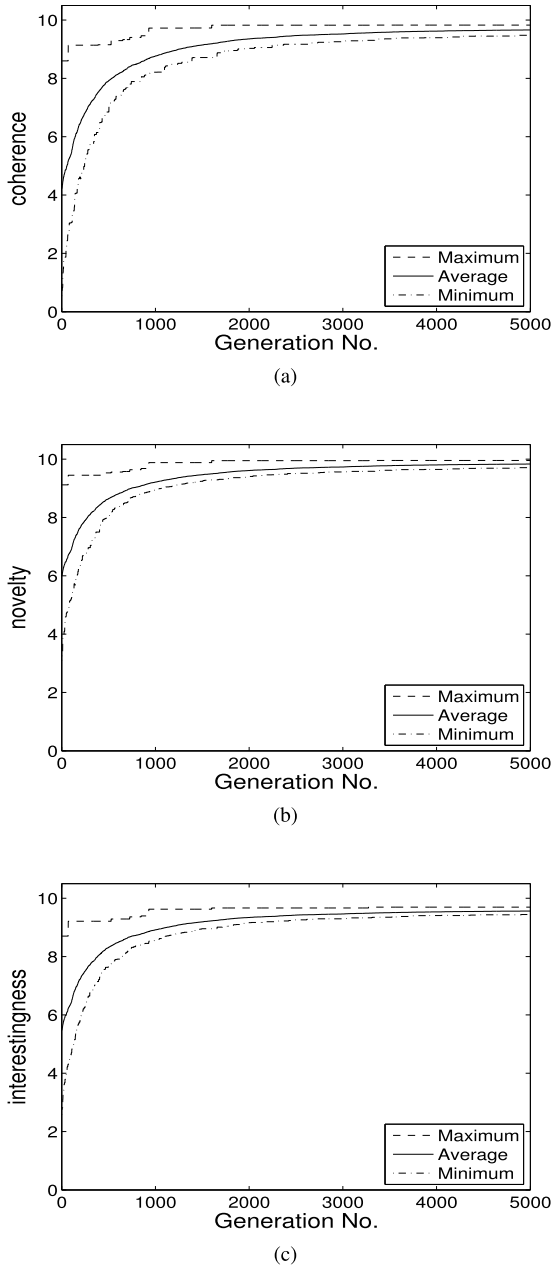

FIGURE 15. Transition of values of subjective metrics during evolutionary process: plots of the minimum, average and maximum values of (a) coherence, (b) novelty, and (c) interestingness among the story narration individuals in the population of each generation.

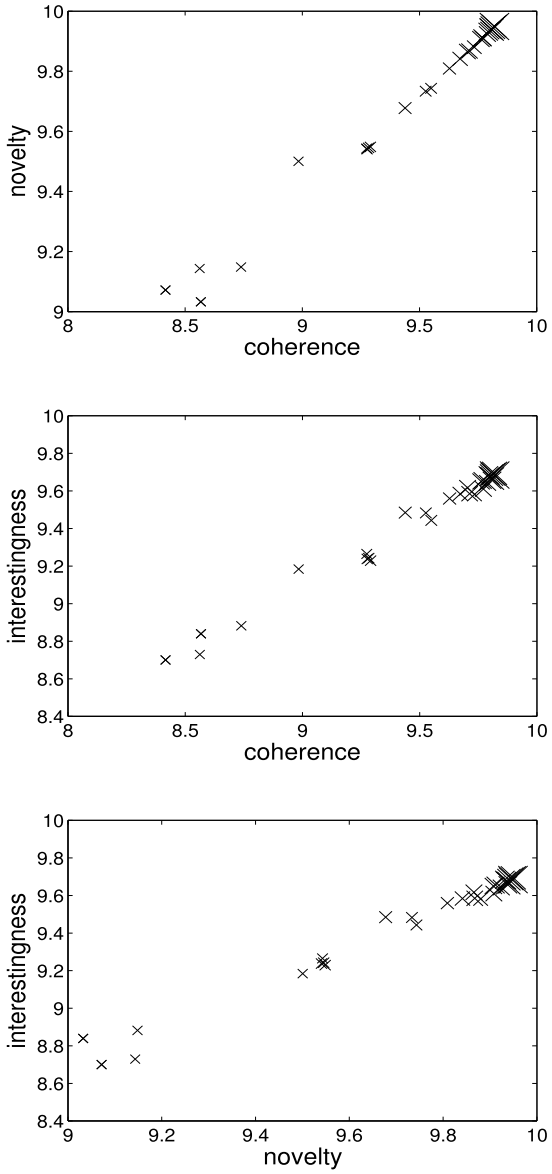

FIGURE 16. Transition of non-dominated fronts in subjective metrics dimension during evolutionary process: non-dominated fronts in subsequent generations are denoted by cross points with bigger sizes.

from the storytelling systems due to the complexity of language required as story length increases.

Second, as it is difficult to quantify the quality of a story using a uniform set of objective metrics, the ultimate judge for the quality of a story is still a human-being.

However, one difficulty lies in the need for a large number of human subjects to evaluate the stories. Humans are diverse in their taste. Moreover, conducting comparable human-based experiments is challenging considering environmental and human expertise difference.

Consequently, many studies have focused on realising the authentic features of story or "storiness" in their specific domain of interest rather than evaluating the stories produced by their systems. A few pieces of work have introduced a pure ex-post-facto story evaluation stage to verify their story generation methodology, including MINSTREL [58] and Picture Books [59], in which small-scale human-based story evaluation experiments were carried out, and Prevoyant [12], in which 54 human participants evaluated stories with different degrees of surprise due to flashback and foreshadowing.

A human-based verification experiment is carried out in this section to verify our proposed human-guided evolutionary story narration approach in its capability to collect

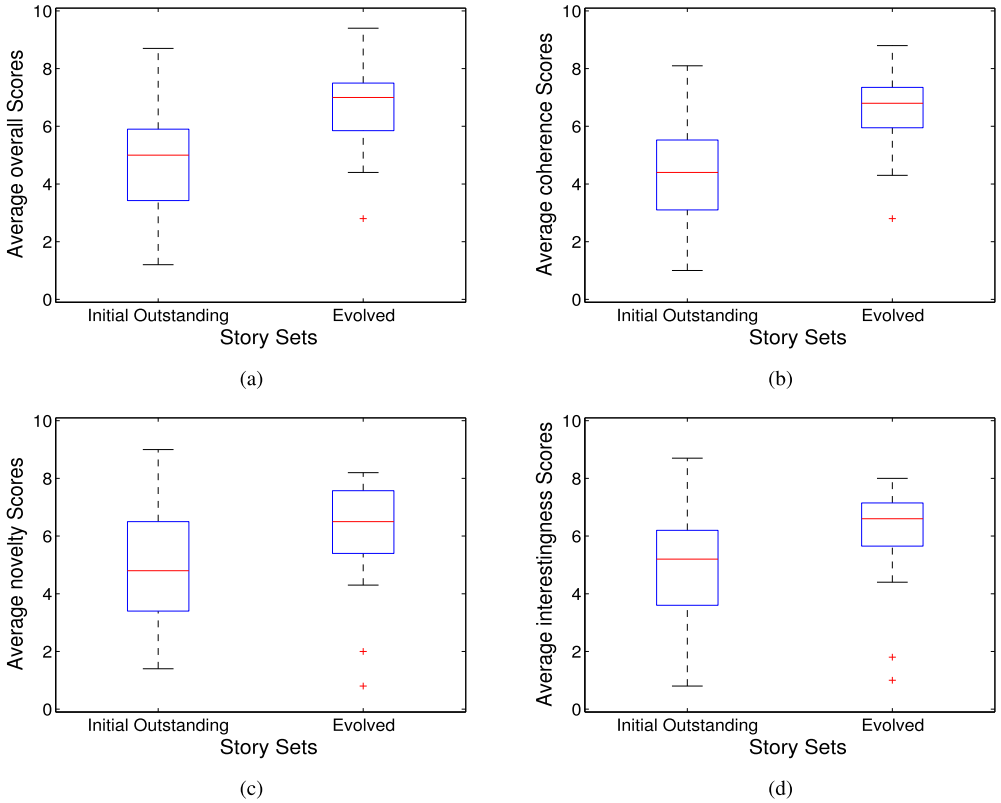

FIGURE 17. Observed subjective metrics distributions of outstanding initial and evolved story narrations: (a) overall, (b) coherence, (c) novelty, and (d) interestingness.

good story narrations from human subjective perspectives. Randomly generated story narrations and evolved ones are evaluated and compared by 31 human participants.

A. EXPERIMENTAL DESIGN

1) HUMAN PARTICIPANT SAMPLE

The same 31 of 42 participants (11 apologised and were absent due to conflicting schedule) as the previous human-based evaluation experiment discussed in Section V are involved.

2) STORY NARRATION SAMPLE

We need to allow a certain degree of variance in the story narration sample to alleviate the granularity and inconsistency problem in IEC [30]. On one hand, humans subjective evaluation for two similar story narrations (i.e. narrations with minor changes in the feature parameter space) may fail to show proper granularity considering the minor changes in their mental picture built in the psychological space. On the other, humans subjective evaluation for same story narration may fluctuate according to time. Therefore, using the evolutionary story narration process proposed in Section VII, the story narration sample is obtained in the following steps:

Firstly, the evolutionary processes are run for 20 times using 20 different seeds and the information of the story narration individuals in the following data set (1) and (2) collected.

(1) Non-dominated individuals in the initial populations.

(2) Non-dominated individuals in the final populations.

Then, 11 individuals are chosen which comprise the story narration sample: 5 with comparatively varied values of objective metrics selected from dataset (1); 5 with varied values of objective metrics from dataset (2); and one story narration individual corresponding to the original CaveLand story.

The text-form of the above 11 individuals in the final story narration sample are presented to the 31 participants in the above human participant sample to evaluate. Their scores for the subjective story metrics of each story narration — the scores for 341 stories (11 story narration individuals $\times$ 31 participants) — are collected and analysed.

B. RESULTS AND ANALYSIS

1) FITNESS OF SURROGATE MODEL FOR HUMAN STORY EVALUATION

Table 4 provides the discrepancy of human story evaluation prediction using the surrogate model based on the following two data sets: data under the "Original Data" title which is calculated from the values of the approximated subjective metrics and the original scores collected from the human participants in the human-based verification experiment; and data under the "Normalised Data" title is obtained from the normalised scores calculated using Equation (5).

TABLE 4. Discrepancy of human story evaluation prediction using surrogate model for human story evaluation.

Subjective Metrics	Original	Normalised Data
overall	3.13	1.97
coherence	3.43	2.45
novelty	3.76	1.91
interestingness	3.42	1.92

Drawn from the positive values of discrepancy shown in Table 4, the surrogate model tends to overestimate human participants' subjective evaluations on the produced story narrations. The lower discrepancy from the normalised data implies that the introduction of a reference story to successive human-based story evaluation experiments may be necessary.

R^2 of the surrogate mode indicates better fitness than the surrogate model in the preliminary implementation discussed in a previous work [27], which shows that the surrogate model explains 70.25% of the variability of the observed evaluation of the overall story quality, 62.59% for coherence, 51.21% for novelty, and 60.44% for interestingness.

TABLE 5. Proportions of human participants favouring evolved story narrations in a human-based verification experiment.

Subjective Metrics	Proportion of Participants
overall	30 of 31 (96.77%)
coherence	31 of 31 (100%)
novelty	25 of 31 (80.65%)
interestingness	27 of 31 (87.10%)

2) EFFECTIVENESS IN GENERATING IMPROVED STORIES

Table 5 presents the proportion of human participants that are in favour of the evolved story narrations (those in dataset (2) in the above story narration sample) rather than the outstanding initial story narrations (in dataset (1)). It shows that the majority of the human participants regard evolved story narrations as those with improved quality compared with the already outstanding randomly generated ones. In particular, 100%human participants agree that the evolutionary process can produce coherent story narrations. Also, a minority of human participants, 19.35% for the novelty and 12.9% for interestingness, still prefer the unevolved stories in terms of their novelty and interestingness.

Fig.17 compares the average scores of the subjective metrics provided by the human participants to the 5 outstanding initial story narrations and the 5 evolved ones in the story narration sample. It further verifies the effectiveness of the evolutionary story narration process in its capability to collect story narrations with overall improved quality.

IX. CONCLUSIONS AND FUTURE WORK

Existing automatic storytelling approaches possess several limitations, ranging from the possible lack of creativity, coherence and interestingness to the absence of human readers' assessment in the generated stories. We have shown in an automatic story narration application that the ability to stochastically evolve a population of stories using interactive evolutionary computation (IEC) techniques is a possible solution to these problems. A human-guided evolutionary story narration approach is proposed and discussed. This extends existing work by transforming a story narration problem into a classical permutation problem in EC, and by devising a human-based story evaluation schema that can incorporate diversified human tastes. The results of the conducted human-based verification experiment demonstrate that this approach is effective in evolving better story narrations from randomly generated ones as assessed by 31 human participants.

The problems and possible future work which can extend this paper are as follows. Firstly, human interaction was only involved once in the evolutionary process in order to minimise human input as far as possible. In future work, human interaction in story evaluation can adopt a mid-point at which a human interacts with story evolution every now and then to progressively adapt the surrogate model of human evaluation which may produce even better-quality stories at the end. Also, the generated story narrations are based on the same story content. Future work involves defining a story grammar from different stories in our domain of interest in which the proposed story parsing method may serve as a pragmatic tool, so that a story structure on the plot level can be manipulated by the IEC process to evolve stories that are thematically different. Besides, a length limit to each story presented to human evaluators seems appropriate considering that humans need to understand a story to give proper evaluations and the evaluations affect the precision of the surrogate model which guides the story evolution. We conjecture that the evolution of long and complex stories may be facilitated after conducting event tagging and grouping into chains using our proposed story parsing method, and presenting each chain in a concise and human comprehensible way.

REFERENCES

[1] G. Kelly, *A Theory of Personality: The Psychology of Personality Constructs*. New York, NY, USA: Norton, 1963.
[2] W. Labov and J. Waletzky, "Narrative analysis: Oral versions of personal experience," *J. Narrative Life Hist.*, vol. 7, nos. 1–4, pp. 3–38, 1997.
[3] A. J. Cowell *et al.*, "Understanding the dynamics of collaborative multi-party discourse," *Inf. Vis.*, vol. 5, no. 4, pp. 250–259, 2006.
[4] H.-M. Chang and V.-W. Soo, "Planning-based narrative generation in simulated game universes," *IEEE Trans. Comput. Intell. AI Games*, vol. 1, no. 3, pp. 200–213, Sep. 2009.
[5] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, "Story plot generation based on CBR," *Knowl.-Based Syst.*, vol. 18, nos. 4–5, pp. 235–242, Aug. 2005.
[6] Y.-T. C. Yang and W.-C. I. Wu, "Digital storytelling for enhancing student academic achievement, critical thinking, and learning motivation: A year-long experimental study," *Comput. Edu.*, vol. 59, no. 2, pp. 339–352, 2012.
[7] H. A. Abbass, S. Alam, and A. Bender, "Mebra: Multiobjective evolutionary-based risk assessment," *IEEE Comput. Intell. Mag.*, vol. 4, no. 3, pp. 29–36, Aug. 2009.
[8] V. Bui, A. Bender, and H. Abbass, "An expressive GL-2 grammar for representing story-like scenarios," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2012, pp. 1–8.

[9] I. Swartjes, "The plot thickens: Bringing structure and meaning into automated story generation," Ph.D. dissertation, Dept. Elect. Eng., Math. Comput. Sci., Univ. Twente, Enschede, The Netherlands, Mar. 2006.

[10] N. Szilas, "Idtension: A narrative engine for interactive drama," in *Proc. Technol. Interact. Digit. Storytelling Entertainment (TIDSE)*, vol. 3. 2003, pp. 187–203.

[11] J. Porteous, M. Cavazza, and F. Charles, "Applying planning to interactive storytelling: Narrative control using state constraints," *ACM Trans Intell. Syst. Technol. (TIST)*, vol. 1, no. 2, pp. 111–130, 2010.

[12] B.-C. Bae and R. M. Young, "A computational model of narrative generation for surprise arousal," *IEEE Trans. Comput. Intell. AI Games*, vol. 6, no. 2, pp. 131–143, Jun. 2014.

[13] J.-P. Kelly, A. Botea, and S. Koenig, "Offline planning with hierarchical task networks in video games," in *Proc. 4th Artif. Intell. Interact. Digit. Entertainment (AIIDE) Conf.*, 2008, pp. 60–65.

[14] Y. Cai, C. Miao, A.-H. Tan, and Z. Shen, "Context modeling with evolutionary fuzzy cognitive map in interactive storytelling," in *Proc. IEEE World Congre. Comput. Intell.*, Jun. 2008, pp. 2320–2325.

[15] Y.-G. Cheong and R. M. Young, "A computational model of narrative generation for suspense," in *Proc. AAAI*, 2006, pp. 1906–1907.

[16] M. Mateas and A. Stern, "Writing façade: A case study in procedural authorship," in *Proc. 2nd Person: Role-Playing Story Games Playable Media*, 2007, pp. 183–208.

[17] M. Garber-Barron and M. Si, "Adaptive storytelling through user understanding," in *Proc. 9th Artif. Intell. Interact. Digit. Entertainment Conf.*, 2013, pp. 128–134.

[18] J. Niehaus and R. M. Young, "Cognitive models of discourse comprehension for narrative generation," *Literary Linguistic Comput.*, vol. 29, no. 4, pp. 561–582, 2014.

[19] H.-Y. Wu, M. Young, and M. Christie, "A cognitive-based model of flashbacks for computational narratives," in *Proc. 12th AAAI Conf. Artif. Intell. Interact. Digit. Entertainment*, 2016, pp. 239–245. [Online]. Available: http://aaai.org/ocs/index.php/AIIDE/AIIDE16/paper/view/13988

[20] K. Wang, E. Petraki, and H. Abbass, *Evolving Narrations of Strategic Defence and Security Scenarios for Computational Scenario Planning*. Cham, Switzerland: Springer, 2016, pp. 635–661.

[21] M. Theune, S. Faas, A. Nijholt, and D. Heylen, "The virtual storyteller: Story creation by intelligent agents," in *Proc. Technol. Interact. Digit. Storytelling Entertainment (TIDSE) Conf.*, 2003, pp. 204–215.

[22] H. Manurung, "An evolutionary algorithm approach to poetry generation," Ph.D. dissertation, School Inform., Univ. Edinburgh, Edinburgh, U.K., 2004.

[23] S. Giannatos, Y.-G. Cheong, M. Nelson, and G. N. Yannakakis, "Generating narrative action schemas for suspense," in *Proc. AAAI Conf. Artif. Intell. Interact. Digit. Entertainment*, 2012, pp. 8–13.

[24] V. Bui, H. Abbbass, and A. Bender, "Evolving stories: Grammar evolution for automatic plot generation," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, 2010, pp. 1–8.

[25] K. Wang, V. Bui, and H. Abbass, "Evolving stories: Tree adjoining grammar guided genetic programming for complex plot generation," in *Simulated Evolution and Learning* (Lecture Notes in Computer Science), vol. 6457, K. Deb *et al.*, Eds. Berlin, Germany: Springer, 2010, pp. 135–145.

[26] K. Wang, V. Bui, E. Petraki, and H. A. Abbass, "From subjective to objective metrics for evolutionary story narration using event permutations," in *Simulated Evolution and Learning* (Lecture Notes in Computer Science), vol. 7673, L. Bui, Y. Ong, N. Hoai, H. Ishibuchi, and P. Suganthan, Eds. Berlin, Germany: Springer, 2012, pp. 400–409.

[27] K. Wang, V. Bui, E. Petraki, and H. A. Abbass, "Evolving story narrative using surrogate models of human judgement," in *Robot Intelligence Technology and Applications* (Advances in Intelligent Systems and Computing), vol. 208, J.-H. Kim, E. T. Matson, H. Myung, and P. Xu, Eds. Berlin, Germany: Springer, 2013, pp. 653–661.

[28] T. Ong and J. J. Leggett, "A genetic algorithm approach to interactive narrative generation," in *Proc. 15th ACM Conf. Hypertext Hypermedia*, 2004, pp. 181–182.

[29] R. P. Y. Pérez and M. Sharples, "Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA," *Knowl.-Based Syst.*, vol. 17, no. 1, pp. 15–29, 2004.

[30] H. Takagi, "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation," *Proc. IEEE*, vol. 89, no. 9, pp. 1275–1296, Sep. 2001.

[31] X. Sun, D. Gong, Y. Jin, and S. Chen, "A new surrogate-assisted interactive genetic algorithm with weighted semisupervised learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 685–698, Apr. 2013.

[32] J. Branke, S. Greco, R. Słowiński, and P. Zielniewicz, "Learning value functions in interactive evolutionary multiobjective optimization," *IEEE Trans. Evol. Comput.*, vol. 19, no. 1, pp. 88–102, Feb. 2015.

[33] M.-L. Ryan, "Toward a definition of narrative," in *Cambridge Companion to Narrative* (Cambridge Companions to Literature), D. Herman, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2007, pp. 22–35.

[34] Oxford. (2013). [Online]. Available: http://oxforddictionaries.com/

[35] R. Saurı, L. Goldberg, M. Verhagen, and J. Pustejovsky, "Annotating events in English. TimeML annotation guidelines," in *Proc. 4th Int. Workshop Semantic Eval. (SemEval)*, Prague, Czech, Jun. 2009.

[36] G. Puscasu and V. B. Mititelu, "Annotation of WordNet verbs with TimeML event classes," in *Proc. 6th Int. Conf. Lang. Resour. Eval. (LREC)*, Marrakech, Morocco, May 2008, pp. 2793–2800. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/

[37] Dictionary.com (2013). [Online]. Available: http://dictionary.reference.com/

[38] G. N. Leech, G. Leech, and J. Svartvik, *A Communicative Grammar of English*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[39] T. Trabasso, P. van den Broek, and S. Y. Suh, "Logical necessity and transitivity of causal relations in stories," *Discourse Processes*, vol. 12, no. 1, pp. 1–25, 1989.

[40] B. Boguraev and R. K. Ando, "TimeML-compliant text analysis for temporal reasoning," in *Proc. 19th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 5. 2005, pp. 997–1003.

[41] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics*, 2003, pp. 423–430.

[42] V. Propp, *Morphology of the Folktale*, vol. 9. Austin, TX, USA: Univ. Texas Press, 1973.

[43] M. Theune, S. Rensen, R. den Akker, D. Heylen, and A. Nijholt, "Emotional characters for automatic plot creation," in *Technologies for Interactive Digital Storytelling and Entertainment*. Berlin, Germany: Springer, 2004, pp. 95–100.

[44] J. Truby, *The Anatomy of Story: 22 Steps to Becoming a Master Storyteller*. London, U.K.: Faber and Faber, 2007.

[45] M. Turim, "Flashbacks in film: Memory and history," *J. Aesthetics Art Criticism*, vol. 49, no. 2, p. 191, 1991.

[46] J. Lu, Y. Zhu, X. Zeng, L. Koehl, J. Ma, and G. Zhang, "A linguistic multicriteria group decision support system for fabric hand evaluation," *Fuzzy Optim. Decision Making*, vol. 8, no. 4, pp. 395–413, 2009.

[47] A. Karmiloff-Smith, "Language and cognitive processes from a developmental perspective," *Language Cognit. Processes*, vol. 1, no. 1, pp. 61–85, 1985.

[48] R. M. Young, "Computational creativity in narrative generation: Utility and novelty based on models of story comprehension," in *Proc. AAAI Spring Symp.*, Stanford, CA, USA, 2008, pp. 149–155.

[49] M. A. Boden, "Creativity and artificial intelligence," *Artif. Intell.*, vol. 103, nos. 1–2, pp. 347–356, 1998.

[50] N. Campion, D. Martins, and A. Wilhelm, "Contradictions and predictions: Two sources of uncertainty that raise the cognitive interest of readers," *Discourse Processes*, vol. 46, no. 4, pp. 341–368, 2009.

[51] T. Bridgeman, "Time and space," in *Cambridge Companion to Narrative* (Cambridge Companions to Literature), D. Herman, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2007, pp. 52–65.

[52] S. Weisberg, *Applied Linear Regression*, vol. 528. Hoboken, NJ, USA: Wiley, 2005.

[53] T. Tusar and B. Filipic, "Visualization of Pareto front approximations in evolutionary multiobjective optimization: A critical review and the prosection method," *IEEE Trans. Evol. Comput.*, vol. 19, no. 2, pp. 225–245, Apr. 2015.

[54] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[55] Z. Michalewicz and D. B. Fogel, *How to Solve it: Modern Heuristics*. New York, NY, USA: Springer-Verlag, 2004.

[56] O. Abdoun and J. Abouchabaka, "A comparative study of adaptive crossover operators for genetic algorithms to resolve the traveling salesman problem," *Int. J. Comput. Appl.*, vol. 31, no. 11, pp. 49–57, 2011.

[57] K. Deb, "Multi-objective optimization," in *Search Methodologies*. New York, NY, USA: Springer, 2014, pp. 403–449.

[58] S. R. Turner, "Minstrel: A computer model of creativity and storytelling," Ph.D. dissertation, Dept. Comput. Sci., Univ. California, Los Angeles, CA, USA, 1993.
[59] E. C. Ong, "A commonsense knowledge base for generating children's stories," in *Proc. AAAI Fall Symp. Ser. Common Sense Knowl.*, 2010, pp. 82–87.

KUN WANG received the B.Sc. and M.Sc. degrees from the Ocean University of China, Qingdao, China, in 2006 and 2009, respectively, and the Ph.D. degree from the University of New South Wales, Canberra, ACT, Australia, in 2013. She is currently a Lecturer in control theory and control engineering with the Ocean University of China. Her research interests include automatic storytelling, scenario planning, and evolutionary computation.

VINH BUI received the B.E. degree from the Hanoi University of Technology, and the M.Sc. and Ph.D. degrees from the University of New South Wales, Australia, in 2001 and 2008, respectively. He is currently with the School of Business and Tourism, Southern Cross University, Australia. His research interests include scenario planning, distributed systems, computer and communication networks, and network tomography.

ELENI PETRAKI is currently an Assistant Professor with the University of Canberra, Australia. Her research interests include language teaching and learning in general. One of her research fields includes discourse analysis and intercultural communication and specifically the strategies employed by language users in everyday communication and in the language classroom. Her research interests also include the way grammar is integrated in the language classroom and the implications for Language Teacher education and training.

HUSSEIN A. ABBASS is currently a Professor of information technology with the University of New South Wales, Australian Defence Force Academy, Canberra, Australia. His current research contributes to trusted autonomy with an aim to design next generation trusted artificial intelligence systems that seamlessly integrate humans and machines, artificial intelligence, big data, cognitive science, operations research, and robotics. He is a fellow of the Australian Computer Society, the Operational Research Society, FORS, U.K., the Australian Institute of Management, and the Vice-President for Technical Activities (2016–2017) of the IEEE Computational Intelligence Society. He is an Associate Editor of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, *IEEE Computational Intelligence Magazine*, and four other journals.